An Algorithmic Framework for Positive Action

OLIVER THOMAS, Predictive Analytics Lab – PAL, University of Sussex, UK
MIRI ZILKA, Predictive Analytics Lab – PAL, University of Sussex, UK
ADRIAN WELLER, University of Cambridge & The Alan Turing Institute, UK
NOVI QUADRIANTO, Predictive Analytics Lab – PAL, University of Sussex, UK

Positive action is defined within anti-discrimination legislation as voluntary, legal action taken to address an imbalance of opportunity affecting individuals belonging to under-represented groups. Within this theme, we propose a novel algorithmic fairness framework to advance equal representation while respecting anti-discrimination legislation and equal-treatment rights. We use a counterfactual fairness approach to assign one of three outcomes to each candidate: accept; reject; or flagged as a positive action candidate.

CCS Concepts: • **Computing methodologies** → **Machine learning**; *Causal reasoning and diagnostics.*

Additional Key Words and Phrases: Fair Machine Learning, Auditing, Causal Inference

ACM Reference Format:
Oliver Thomas, Miri Zilka, Adrian Weller, and Novi Quadrianto. 2021. An Algorithmic Framework for Positive Action. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21), October 5–9, 2021, –, NY, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3465416.3483303

1 INTRODUCTION

Allocating resources such as jobs or university placements among individuals requires evaluating their suitability for the role. We want to ensure the selection process is fair and that positive outcomes are fairly distributed within the population. Machine Learning (ML) systems are increasingly being used to inform, support, or even directly make decisions within consequential domains, affecting millions of lives [2]. It is therefore necessary to consider how the notions of fair process and fair outcomes translate into algorithmic decision support frameworks [43, 45].

According to anti-discrimination legislation in the E.U., U.S. and U.K., among others, a fair selection process requires equal treatment in the sense that protected attributes, for example, gender and race, are not to be considered within the decision making process without a good reason [3, 9]. Simply ignoring the protected attributes within an algorithmic approach, however, guarantees neither fair process nor fair outcome [15, 33, 40, 43, 45].

Decision support algorithms are commonly trained on a dataset of past decisions. The resulting algorithms, however, may disproportionately predict positive outcomes in favour of the majority[1] over historically under-represented groups [19, 20]. These statistical disparities could arise from two mechanisms: (i) unequal treatment; or (ii) equal treatment, when the status quo in the environment itself is not neutral. The former occurs when the data contains discriminatory past decisions. The latter, when historically under-represented groups struggle to compete with the majority under

[1]The group enjoying an advantage is not always the majority. We aim for clarity of exposition when referring to the over-represented group as the majority.

Manuscript submitted to ACM

a standard *equal-treatment selection process* — a selection process that is 'blind' to the applicant's protected attribute. Often, the statistical disparity in the training data, and as a result, in the model's prediction, is a combination of both.

Enforcing *Demographic Parity* (DP) – an equal fraction of positive outcomes across subgroups – is usually impractical. In addition to hindering the accuracy of the model's predictions, this approach does not typically align with anti-discrimination legislation. A common alternative is to impose an algorithmic fairness constraint that better aligns with the notion of equal treatment, maintaining a disparity in the positive outcome[2] rates [14, 43].

Anti-discrimination legislation acknowledges the need to bridge the gap between equal treatment and equal representation. The Equality Act 2010 (UK) defines *positive action* as 'lawful measures taken to encourage and train people from under-represented groups to help them overcome disadvantages in competing with other applicants'.[3] Examples of positive action include, but are not limited to: additional training opportunities and mentoring programs available to an under-represented group, targeted advertising, outreach, networking and bursaries. For example, Target Oxbridge is a free, UK based programme that 'aims to help black African and Caribbean students and students of mixed race with black African and Caribbean heritage increase their chances of getting into the Universities of Oxford or Cambridge' [34]. Policies designed to meet the specific needs of under-represented groups may also be considered as positive action. The European Research Council introduced automatic extensions of eligibility only for women with children when applying for grants.[4] The action taken is required to be 'proportionate' to both the extent and longevity of the under-representation, and to the barriers experienced by the under-represented group.

We argue that incorporating the notion of positive action within decision support algorithms can advance the use of positive action measures, promoting equal representation while respecting anti-discrimination legislation and equal-treatment rights. In this work, we propose a novel algorithmic fairness framework to identify *positive action candidates* — individuals who would be rejected under a standard equal-treatment selection process because of earlier disadvantage experienced due to their under-represented group membership.

2 BACKGROUND

2.1 Definitions

In this work we discuss subgroups with respect to *protected attributes* – characteristics that, by law, must not be the basis for discrimination. These include, but are not limited to, race, gender, age, religion and disability. We define a protected subgroup as an under-represented group separated from the majority by the value of one or more protected attributes. For example, women in the engineering profession are under-represented when compared to their representation within the population. In the context of a decision support system, we may observe a statistical disparity — a disproportionate positive outcome (hiring, admissions) rate — in favour of the majority, compared to a protected subgroup. This can be as a result of the model being trained on past discriminatory decisions, but can also be the result of a genuine statistical difference in the input features (grades, qualifications). In this work, we define bias to be a mechanism by which statistical disparity between a protected subgroup and the majority is created or exacerbated. Bias within the decision making process will affect the decision outcome. Bias that occurred earlier may affect the features. To expand the discussion on bias, it is useful to refer to the framework presented by Friedler et al. [11], which defines three spaces

[2]Throughout this paper, we use the terms 'accepted', 'successful' and 'positive' when referring to outcomes interchangeably.
[3]European legislation defines positive action similarly. In the US, similar measures can be employed under affirmative action, however, the definitions do not completely overlap.
[4]These measures are included in the European Research Council's Gender Equality Plan for 2021-2027.

– the *construct space*, *observed space* and *decision space* – and uses the mappings between them to formalise several definitions of bias.

The *construct space* represents the *'ground truth'* – an unobserved space that correctly captures differences between individuals with respect to a task; the *observed space* represents the measurable features for consideration, and the *decision space* represents the outcome [11]. For example, intelligence resides in the construct space, measured IQ resides in the observed space, and acceptance or rejection from the International Mensa Club resides in the decision space.

The observed space allows us to estimate the construct space, but we are required to make assumptions regarding the mapping between spaces. Friedler et al. [11] refers to these assumptions as *worldviews*, highlighting two common worldviews, WAE ("we're all equal") and WYSIWYG ("what you see is what you get"), that are often in tension with each other. WAE assumes that any disparity between subgroups in the observed space is due to structural bias — an incorrect mapping between the construct and the observed space. WYSIWYG, on the other hand, allows for a disparity between protected subgroups, assuming the observed discrepancies are a true reflection of disparities in the construct space. In this work we adopt a 'hybrid' worldview, described in Section 3.1.1, that allows a version of both worldviews to co-exist.

To better understand the potential for statistical disparity in the data, we discuss specific types of bias. *Sample selection bias* originates from training on a non-representative sample of the population [41]. *Label bias* occurs when the dataset contains past discriminatory decisions [17, 44]. Mitigation efforts that consider selection bias [1, 20], or label bias [7, 17, 25] independently are available. Bias can also be introduced from outside the environment we can control, i.e. outside the training population, measurements, and learning algorithm. Our proposed framework aims to acknowledge and mitigate a broader range of biases. This includes bias that cannot normally be mitigated by an automated rejection / acceptance model while respecting anti-discrimination legislation and the right for equal treatment.

2.2 Counterfactual Modelling

To identify positive action candidates, we take a counterfactual approach. A counterfactual outcome is a hypothetical outcome for a scenario that is identical in all respects except for a specific, well-defined change and its causal consequences [16, 29]. In the context of this work, we focus on counterfactual scenarios with respect to a change in a protected attribute, and distinguish between two types of counterfactual questions:

Question 1: Would the outcome change if *only* the protected attribute was different?

Question 2: Would the outcome change if the protected attributed *and its causal consequences* were different? For example, if a female applicant is not invited for a job interview, we can ask the following two questions: if her CV was identical, but the application *appeared* to be from a male applicant, would she be invited to interview?[5] If she had been **born** male, experienced life as a male, and then applied for the same job, would she have been invited for an interview? The second counterfactual question is critical to our approach, as it is used to identify positive action candidates. We use the first counterfactual question to detect and mitigate label bias.

To evaluate counterfactual outcomes, ideally, we would rely on a Structural Causal Model (SCM) — a graphical model whose vertices represent features and whose edges represent the *causal* pathway between them [32]. However, a complete structural model is challenging to obtain — they are application specific and require specification by domain experts. In practice, we can find two as-close-as-possible individuals (differing by the protected attribute) within the data (e.g. [35, 36]) or by creating the counterfactual representations using an adversarial learning model (e.g. [12, 24, 26, 39]).

[5] An experiment by Bertrand and Mullainathan [4] looked at exactly this question.

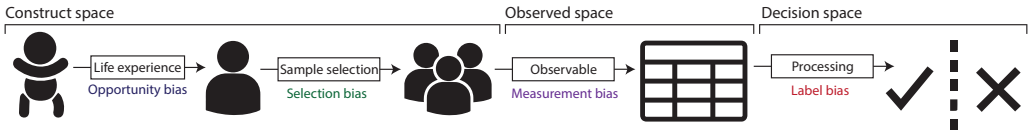

Fig. 1. A 'hybrid' worldview showing biases potentially introduced at each step of a timeline leading up to a decision. Aptitude is characterised by the infant at the beginning of the timeline and is assumed to be independent of all protected attributes, aligning with the WAE worldview. By the point of observation however, the construct space might have altered and our 'hybrid' worldview allows for disparity between subgroups, aligning with the WYSIWYG worldview. Opportunity bias, selection bias, measurement bias and label bias can introduce or further aggravate the disparity between the protected subgroup and the majority.

In this paper, we employ the latter approach. To place our Positive Action framework within the context of other existing works, we include a comparison with related works in Appendix A.

3 APPROACH

We propose a novel algorithmic fairness framework for advancing equal representation while respecting anti-discrimination legislation and the right for equal treatment. We identify *positive action candidates*: roughly speaking, individuals who would be rejected under a standard equal-treatment selection process because of earlier disadvantage experienced due to their under-represented group membership. More precisely, we use counterfactual methods to assign each applicant to one of three groups:

(1) Successful applicants, and applicants from under-represented groups who were unsuccessful, but would have been successful if they had a different set of protected attributes (without considering causal consequences, i.e. Question 1 above) – these are accepted;
(2) Unsuccessful applicants from under-represented groups who are not in (a), and for whom there exists a different set of protected attributes which would have caused them to be successful (considering causal consequences, i.e. Question 2 above) – these are flagged as positive action candidates; and
(3) Everyone else – these are rejected.

Note that for applicants from the majority group, our approach always leaves the outcome unchanged as either accepted or rejected.

3.1 Positive Action Framework

3.1.1 Fairness Worldview. Where does positive action fit within technical fairness definitions? If we consider WAE and WYSIWYG, the worldviews discussed in Section 2.1, this notion does not fully align with either worldview. The 'hybrid' worldview adopted in this work is illustrated in Figure 1 and as a graphical model in Figure 3 .

We expand the *construct space* to include *time*, with the observed space representing measurements of the construct space at points in time. Consider a set of measurable features X within the observed space $\mathcal{X}$, where $\mathcal{X}$ represents the space of all potential feature values. Each individual sample $x \in X$ is an approximation to its non-measurable construct-counterpart $\tilde{x} \in \tilde{X}$, giving the decomposition $X \approx \tilde{X} = \alpha \cdot \tilde{X}_{apt} + \beta \cdot \Delta\tilde{X}$ where $\tilde{X}_{apt} \perp S$ and $\Delta\tilde{X} \not\perp S$ with S being the protected attribute, and α and β being non-negative values that sum to 1. In words, we assume an individual's suitability for the task, at the time of measurement, is a combination of their aptitude ($\tilde{X}_{apt}$), a natural born ability, and their experiences over time ($\Delta\tilde{X}$).[6] We further assume that the aptitude component, $\tilde{X}_{apt}$, is independent of any

[6]We make no claims regarding the strength of 'nature' vs. 'nurture'. The framework holds for all potential ratios, including those where $\alpha = 0$ or $\beta = 0$.

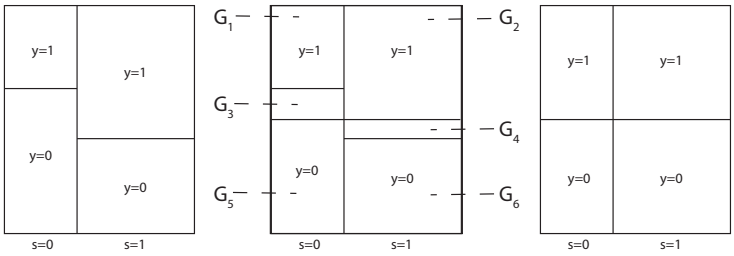

Fig. 2. The accepted ($y = 1$) and rejected ($y = 0$) ratios difference between a protected subgroup ($s = 0$) and the majority ($s = 1$). *Left*: under a standard equal-treatment selection rule (WYSIWYG worldview). *Right*: when demographic parity is enforced (WAE worldview). *Middle*: Overlapping the two worldviews. The population captured by groups G_1, G_2, G_5 and G_6 have consistent outcome across both worldviews. Groups G_3 and G_4 represent individuals that will receive different outcome under the different worldview.

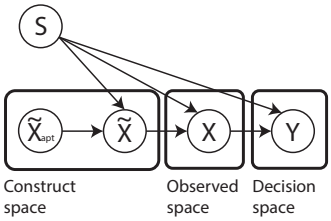

Fig. 3. The effect of a protected attribute S on descendants of $\tilde{X}_{apt}$ throughout a data-generation procedure. $\tilde{X}$ within the construct space, $\mathcal{X}$ within the observed space and $\mathcal{Y}$ within the decision space.

protected attribute, and hence complies with the WAE worldview.[7] The 'life-experience' component $\Delta\tilde{X}$, shifts the aptitude either positively or negatively, and may not be independent of S. $\tilde{X}$ represents the non-observable 'ground truth' at the time of measurement, which could be dependent on S.

3.1.2 Underlying mechanisms and bias. We consider a setting where we observe a statistical disparity between subgroups separated by the value of $\mathcal{S}$, within both the observed space and the decision space. The disparity within the decision space may be worse than the disparity within the observed space. One mechanism that can cause this aggravation is label bias – a direct impact of the protected attribute $\mathcal{S}$ on the outcome $\mathcal{Y}$ due to past discriminatory decisions within the training dataset. To achieve equal treatment, the effects of label bias should be eliminated. The disparity within the observed space can be caused by several mechanisms or their combination: selection bias occurs when the training set contains a non-representative sample of the population; measurement bias occurs when the mapping from the construct space to the observed space isn't as faithful for certain groups or individuals. Furthermore, part of the disparity within the observed space can be a true reflection of a disparity within the construct space itself, at the time of measurement. We assume that the distribution of aptitude $\mathcal{X}_{apt}$ in the construct space is the same across subgroups. While variation in opportunities between individuals is normal, when the imbalance of opportunity affects a protected group more than the majority, it will result in a disparity between the subgroups within the construct space itself. Addressing this imbalance of opportunity is a principal component of positive action and our framework.[8]

3.2 Positive Action Candidates

3.2.1 Quantifying the difference between WAE and WYSIWYG. To quantify the difference between the WAE and the WYSIWYG worldviews we divide the data into six subgroups, as shown in Figure 2. This procedure can be done for any pair of fairness metrics and definitions. We compare positive outcome ratios between an equal-treatment selection rule and demographic parity, metrics associated with WYSIWYG and WAE, respectively. We conceptually overlay the observed data[9] (Figure 2, left) on a representation of the data with demographic parity enforced (Figure 2, right).

[7] We are excluding tasks where success may be strongly correlated with physical attributes – for example, playing professional basketball and height.
[8] We note that this is not an extensive discussion of bias and there are other underlying mechanisms that can lead to a statistical disparity between an under-represented group and the majority.
[9] The outcomes in the observed data are based on a standard equal-treatment selection rule.

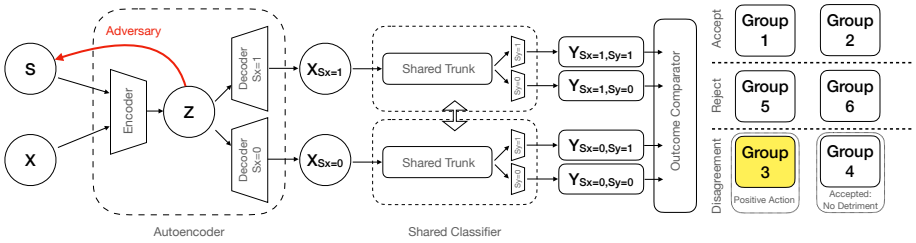

Fig. 4. Diagram illustrating our method. The original representation x is mapped to a representation z that is independent of the protected attribute s. The invariant representation z is then mapped back into both $x_{s_x=0}$ and $x_{s_x=1}$, reintroducing biases associated with each subgroup. Each of those representations is labelled, resulting in four representation in total. The four corresponding predicted outcomes then determine the group classification according to one of three final outcomes: *accept*, *reject*, or *disagreement* which has two outcomes associated. Candidates from under-represented groups that were rejected, but would have received a positive outcome in a counterfactual world are *flagged for positive action*. Candidates from majority s that were flagged for acceptance, but would *not* have received a positive outcome in a counterfactual world remain accepted under a 'no-detriment' policy.

When overlaid, the data can be separated into six subgroups, as shown in Figure 2, middle. Subgroups G_1 and G_2 get a positive outcome under both worldviews. Subgroups G_5 and G_6, get a negative outcome under both worldviews. Subgroups G_3 and G_4, however, represent *a different outcome under the two worldviews*. Subgroup G_3, represents the subgroup that would have received a positive outcome had demographic parity been enforced, and a negative outcome based on the observed data. This subgroup may be interpreted as individuals who would be rejected under a standard equal-treatment selection process because of earlier disadvantage experienced due to their under-represented group membership. We cannot accept these applicants while aligning with anti-discrimination legislation. However, we can highlight them as candidates for positive action — targeted support to help them succeed under an equal-treatment selection process.

3.2.2 Choosing positive action candidates. Demographic parity is a group fairness measure that compares the ratios of positive outcome rates between subgroups. We still need to identify which applicants we want to highlight as positive action candidates. The reader might now consider a straightforward 'baseline' approach of highlighting the top rejected candidates from the under-represented group. This baseline is only applicable when there is a clear way to rank candidates and does not account for two potential issues: measurement bias, and uneven dispersion of disparity amongst the input features. We illustrate these two issues with the following motivating example:

Consider a minority who traditionally sends their children to schools that teach English to a good level but teaches Maths only to a basic level. This minority is under-represented within STEM subjects. To keep this example simple, we consider the application to consist of grades in only two subjects, Maths and English, with equal weight. Blindly taking the best rejected applicants will not spot the applicants who did exceptionally well in Maths, considering the poor education they received in this subject. In our approach, the minority's Maths grade distribution gets re-calibrated to match the majority's distribution, while the distribution of the English grades is left unaffected because there is no disparity with the majority's distribution. Figure 5 illustrates how two applicants would be ranked under our approach compared the baseline of choosing the top rejected candidates. For the majority, the distribution ranges between 0–10 for both English and Maths. For the minority, the English distribution ranges between 1–10 but the Maths distribution only ranges between 1–5. Applicant A's grades are 2 and 9 in Maths and English, respectively. Applicant B's grades are 5 and 6 in Maths and English, respectively. With an equal weight selection rule, they both have an overall score

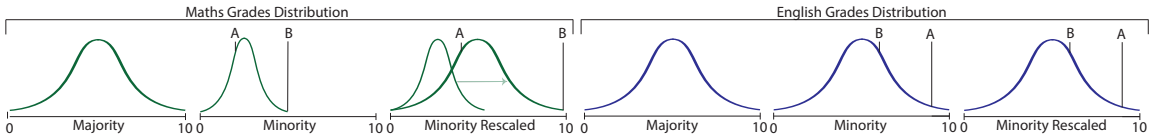

Fig. 5. How our approach to choosing positive action candidates compares to choosing the top rejected candidates from the under-represented group. With an equal weight selection role, Applicants A and B have the same overall score. Re-scaling the minority's Maths grade distribution to match the majority's distribution highlights applicant B as the better positive action candidate.

of 11. When we re-scale the Maths grade distributions of the minority to match the majority's distribution, applicant B is highlighted as the better positive action candidate with an overall score of 16 compared to 13 for applicant B. This re-calibration is only put into effect when populating the positive action candidates group. When applicants are considered for acceptance, the features are taken as they are. In the case of this example, we may not be able to accept applicant B, but they are flagged as a positive action candidate — a Maths foundation course, for example, is likely to allow them to successfully compete in a subsequent selection process.

3.3 Implementation

3.3.1 Building a group classifier. To identify which candidates may benefit most from positive action we use a two-step approach following the scheme in Figure 4. Our aim is to produce, with respect to a protected attribute, both counterfactual samples, and counterfactual decisions. The first accounts for differences in the features. The latter accounts for decisions that are potentially discriminatory.

We follow a common approach from fair representation literature – to make a representation of the data that, as best possible, is invariant to S. First, we train an adversarial autoencoder model that maps the observed data point x from the dataset X into a latent representation $z \in \mathcal{Z}$ (where $\mathcal{Z} = \mathbb{R}^{N_z}$), that is independent of the protected attribute, $s \in \mathcal{S}$, where $\mathcal{S}$ is the set of possible protected attribute values. For example, $\mathcal{S} = \{0, 1\}$ if the protected attribute is binary. From a latent value z, two mirror representations can be created, $x_{s_x=0}$ and $x_{s_x=1}$. The variables $x_{s_x=0}$ and $x_{s_x=1}$ are then labelled by concatenating the perceived protected attribute to the covariates, creating four representations in total: $x_{s_x=0,s_y=0}$, $x_{s_x=0,s_y=1}$, $x_{s_x=1,s_y=0}$ and $x_{s_x=1,s_y=1}$. Here, S_y denotes the value of the protected attribute concatenated to the set of covariates, adding a direct path in the data to S. A classifier can use this value directly if it is indeed the basis of a decision, rather than extracting the protected attribute from the remaining features.

For the next step, we train a second model, a shared classifier, to perform predictions from the counterfactual representations. We then feed in the counterfactuals and get a corresponding set of outputs: $y_{s_x=0,s_y=0}$, $y_{s_x=0,s_y=1}$, $y_{s_x=1,s_y=0}$ and $y_{s_x=1,s_x=1}$. Based on these, the Outcome Comparator then sorts the set of original samples X into one of six subgroups G_{1-6}. The full selection rules are presented in Table 1, but we give some intuition:

Groups 1 & 2 ($G_{1,2}$) consist of candidates whose outcomes were either unanimously accepted across all counterfactual inputs (selection rule 1), or differed due to S_y, the concatenated *perceived* protected attribute, changing (selection rules 2 & 8). Unanimous *negative* outcomes for all counterfactual inputs are assigned to groups $G_{5,6}$ (selection rule 9). Lastly, applicants who receive a disagreement amongst the outcomes, i.e. their outcome depends on the value of S_x, are assigned to groups $G_{3,4}$ (selection rules 3-7). Members of group G_4 are accepted as they would by an unconstrained classifier as our positive action approach has no-detriment to the majority. Members of group G_3 are highlighted as *positive action candidates*.

Table 1. Selection rules for mapping from the groups represented in Figure 2 and Figure 4 to a decision. As $s = 0$ represents an disadvantaged group, we identify those in group 3 to be suitable for *positive action*. Combinations not listed are identified and the outcome reverts to the outcome from an unconstrained model.

Selection Rule	s	$y_{s_x=0,s_y=0}$	$y_{s_x=0,s_y=1}$	$y_{s_x=1,s_y=0}$	$y_{s_x=1,s_y=1}$	Subgroup	y	Outcome
1	0 or 1	1	1	1	1	G_1 or G_2	1	Accept
2	0 or 1	0	1	1	1	G_1 or G_2	1	Accept
3	1	0	0	1	1	G_4	1	Accept
4	1	0	0	0	1	G_4	1	Accept
5	1	0	1	0	1	G_4	1	Accept
6	0	0	0	1	1	G_3	2	Positive Action
7	0	0	0	0	1	G_3	2	Positive Action
8	0	0	1	0	1	G_1	1	Accept
9	0 or 1	0	0	0	0	G_5 or G_6	0	Reject

Our model is implemented as two successive neural networks, each representing one of the distinct phases mentioned above.[10] Implementation details can be found in Appendix B.

4 EXPERIMENTS

We first use synthetic data to demonstrate how our approach can be applied to a candidate-filtering task within a biased setting. We consider applicants to a university course in a fictitious world that is inhabited by *blue* and *green* people, such that we take a person's *colour* as the protected attribute. This university course is for a traditionally *blue* profession, rendering the setting potentially biased. The department receives applications from many more promising *blue* candidates than from promising *green* candidates. We then demonstrate our approach on the UCI Adult Income dataset, and use it to highlight potential challenges in a real-world deployment setting.

4.1 Data

Synthetic Data: We define a data-generation procedure for a dataset with binary S-labels and a binary outcome, with 2 imperfect observers of 3 features, making a feature-space $\mathcal{X}$ comprising 6 features. (Full details are in Appendix F.)
UCI Adult Income Data: We evaluate our approach on the UCI Adult Income Dataset [8], which is often used for evaluating fairness-enhancing systems. This dataset comprises 45, 222 samples from the 1994 U.S. census with 14 features including occupation, maximum attained education level and relationship status. Of these 14 features, we reserve the binary `salary` feature as the target label, with `>$50K` being the positive outcome. We consider 3 binary features, individually, as protected attributes: sex (Male/Female), race (White/Not White) and marital status (Married/Not Married). Discussion of these results can be found in Appendix D.

4.2 Evaluation

To evaluate our model in context, we train the following models on the synthetic data: a Demographic Parity Oracle, `DemPar`, enforcing *exact* Demographic Parity; an unconstrained `Logistic Regression` (`LR`) model ; established fair classification models `K & C Reweighting` [20], `Kamishima` [21] and `FairLearn` [1]; and our positive action approach using counterfactual modelling, which we refer to as `PAF` (Positive Action Framework).

[10] Our code is available at https://github.com/predictive-analytics-lab/positive-action-framework.

Table 2. Comparison table for the synthetic data results. All probability-based results are converted to percentages. Our positive action framework model (PAF) captures 97% (see TCP|G row) of the green applicants capable of graduating: they are either accepted or flagged as positive action candidates. This high TCP value is achieved while maintaining low FID.

	Oracle Values	Unconstrained	Fair models	
Metric	DemPar	LR	FairLearn	PAF (ours)
Acceptance|B	23.12 ± 16.31	34.52 ± 0.79	26.65 ± 1.17	35.15 ± 1.36
Acceptance|G	23.13 ± 16.31	5.43 ± 0.52	7.83 ± 0.52	6.04 ± 0.60
TCP|B ↑	60.57 ± 42.80	91.70 ± 1.92	71.51 ± 3.72	92.63 ± 2.08
TCP|G ↑	69.83 ± 6.03	69.84 ± 5.00	53.17 ± 6.02	96.74 ± 1.93
FID ↓	23.14 ± 14.59	3.73 ± 2.94	20.74 ± 2.95	4.16 ± 3.04
Accuracy(Y) ↑	84.50 ± 0.43	98.51 ± 0.18	92.64 ± 0.41	98.31 ± 0.26
Accuracy(G) ↑	79.25 ± 9.48	87.05 ± 0.46	86.16 ± 0.44	86.67 ± 0.47

Table 3. Comparison of the number of samples allocated to each group for a ground truth Perfect Counterfactual (PCF) in comparison to the Learned Counterfactuals of our Positive Action Framework (PAF). This comparison is only possible because we know the ground truth for the synthetic data.

Subgroup	Outcome	PCF	PAF
g_1	1	4.73 ± 0.15	3.01 ± 0.28
g_2	1	4.73 ± 0.09	2.40 ± 0.48
g_3	2	10.99 ± 0.19	13.77 ± 2.08
g_4	1	10.96 ± 0.25	15.27 ± 0.91
g_5	0	34.35 ± 0.26	32.97 ± 2.86
g_6	0	34.23 ± 0.28	32.59 ± 0.86

We define the following metrics: *Acceptance percentage per colour* (Acceptance), *True Capture percentage* (TCP), *False Identification Difference* (FID), *Accuracy*. See Appendix C for full details.

5 RESULTS & DISCUSSION

5.1 Analysing the baseline synthetic data.

Results comparing the models can be found in Table 2 with a full suite of results available in Appendix E.[11] The synthetic data was engineered to demonstrate a biased setting. We briefly describe the underlying disparity in the data (Table 5 – Data): only 4% of the *green* candidates are admitted, in comparison to 36% of the *blue* candidates. We can evaluate the True Capture percentage (TCP): how many candidates with the ability to graduate, are not being rejected. The potential to graduate, $\mathcal{G}$, represents the 'ground truth' potential and correlates directly to $\tilde{\mathcal{X}}$. While for blue applicants the TCP is at a high 94%, the green TCP is only 60%. The False Identification difference (FID) measures how well the data conforms to a notion of equality of opportunity (EqOP) by reporting the difference in selecting suitable applicants that will successfully graduate. A low FID, 4%, means the data corresponds to an approximately EqOP setting: once a candidate is accepted, the likelihood of graduation is nearly the same for both groups.

Demographic Parity Oracle. When enforcing demographic parity on the model, the metrics change substantially. Acceptance percentage is now equal between the *blue* and *green* candidates, but at a cost: the TCP for blue candidates has gone down to 61% and the FID is up to 23%, meaning the likelihood of graduating once accepted is no longer independent of S.

Unconstrained Baselines. Logistic Regression gives similar results to the baseline data. We use it as a baseline model for the comparison with subsequent models.

Fair Baselines. FairLearn achieved the second best acceptance rate for green applicants, after the DP oracle. Similar to the DP oracle, however, the additional green applicants who get accepted are not the ones capable of graduating.

Positive Action Framework Model. Our PAF model shows a minor improvement in the acceptance rate of green applicants, compared to the baseline data. These additional candidates were flagged by our model as falling victim to label bias, i.e. they would have been accepted if they were *perceived* as blue. They are reassigned to 'accept' by the group classifier (selection rule 8 in Table 1). Unlike the DP oracle (DemPar) and FairLearn, the FID remains low at 4%, as these additional

[11] For ease of comparison, we choose to omit the results for K & C Reweighting and Kamishima from the table and only report FairLearn, the model with the highest acceptance percentage for green applicants out of these three models.

accepted green applicants are capable of graduating. A notable success, in comparison to the other models, is the high TCP combined with a low FID. The addition of the positive action candidate outcome increases the percentage of applicants capable of graduating that are not rejected from 53% to 97%. The positive action candidate outcome enables us to not simply reject high-potential candidates from under-represented groups, even if we are not able to accept them under an equal-treatment selection process. Equal treatment and equality of opportunity are both maintained for accepted applicants.

Table 3 shows the breakdown of the outcome groups, in respect to all the candidates, produced by our PAF model. This is compared to the outcome groups of a 'perfect' counterfactual (PCF), the exact counterfactual value obtained from the synthetic data.[12] We can see that the proportion of the candidates assigned to each group is consistent when comparing the PAF outcomes to those obtained using PCF.

6 LIMITATIONS AND INTENDED USE

In this work, we assume we are required to enforce the mapping between the observed and the decision space to be independent of the protected attribute, i.e., we assume it is a requirement to mitigate for label bias (selection rules 2 & 8, Table 1). This is the only bias that is mitigated at the accept / reject level. The inclusion of the positive action candidate outcome and the G_3 subgroup enables us to audit and mitigate, in the form of recommending candidates for positive action, any additional effects that may cause disparity, i.e. selection bias and imbalance of opportunities.

We choose to adopt a no-detriment, or positive-corrective approach. This means that no individual, even if they allegedly benefit from past biased decisions, will be made worse off by the positive action approach. In practice, selection rules can be adapted to suit the context and objectives at hand.

7 CONCLUSION

We present a novel algorithmic fairness framework that builds on the notion of positive action to advance equal representation while respecting anti-discrimination legislation and the right for equal treatment. We aim not to reject high-potential applicants from under-represented groups, even if they cannot yet successfully compete in an equal-treatment selection process against applicants from the majority group. As we are unable to accept them directly, they are highlighted as promising candidates for positive action measures.

Positive action initiatives can already be found in practice and can include outreach activities, targeted training and adaptive policies. Specific positive action measures will be case and context dependent and should be determined by domain experts. Our aim is to demonstrate that machine learning has the potential to help identify those applicants who would benefit from this additional support.

We consider the different mechanisms that can lead to an observed disparity in the rate of positive outcomes between a protected subgroup and the majority. We highlight that, at least in part, this disparity can be due to disadvantages affecting applicants belonging to a protected subgroup, hindering their ability to compete with other applicants.

Our counterfactual implementation achieves our goal: it maintains predictive utility while minimising the rejection of candidates with high potential from the disadvantaged group. We hope this work will form part of a larger, constructive discussion around the role of machine learning in promoting the use and effectiveness of positive action measures.

[12] These values assume a consistent decision rule across all populations.

ACKNOWLEDGMENTS

This research was supported in part by a European Research Council (ERC) Starting Grant for the project "Bayesian Models and Algorithms for Fairness and Transparency", funded under the European Union's Horizon 2020 Framework Programme (grant agreement no. 851538). AW acknowledges support from a Turing AI Fellowship under grant EP/V025379/1, The Alan Turing Institute under EPSRC grant EP/N510129/1 and TU/B/000074, and the Leverhulme Trust via CFI.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholmsmässan, Stockholm Sweden, 60–69. http://proceedings.mlr.press/v80/agarwal18a.html

[2] S. Barocas and A. Selbst. 2016. Big data's disparate impact. *California Law Review* 104, 3 (2016), 671–732.

[3] Jason R Bent. 2019. Is algorithmic affirmative action legal. *THE GEORGETOWN LAW JOURNAL* 108 (2019), 803.

[4] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review* 94, 4 (2004), 991–1013.

[5] Emily Black, Samuel Yeom, and Matt Fredrikson. 2020. FlipTest: Fairness Testing via Optimal Transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 111–121. https://doi.org/10.1145/3351095.3372845

[6] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York University, New York, USA, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[7] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (01 Sep 2010), 277–292. https://doi.org/10.1007/s10618-010-0190-x

[8] Dua Dheeru and Efi Karra Taniskidou. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) *(ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[10] J. R. Foulds, R. Islam, K. Keya, and S. Pan. 2020. An Intersectional Definition of Fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE Computer Society, Los Alamitos, CA, USA, 1918–1921. https://doi.org/10.1109/ICDE48307.2020.00203

[11] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (Im)Possibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making. *Commun. ACM* 64, 4 (March 2021), 136–143. https://doi.org/10.1145/3433949

[12] Karan Goel, Albert Gu, Yixuan Li, and Christopher Ré. 2021. Model Patching: Closing the Subgroup Performance Gap with Data Augmentation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, Virtual Event, Austria, 32. https://openreview.net/forum?id=9YlaeLfuhJF

[13] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander J Smola. 2007. A Kernel Approach to Comparing Distributions, In Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada. *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence (AAAI-07)* 22, 1637–1641. http://www.aaai.org/Library/AAAI/2007/aaai07-262.php

[14] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates Inc., Red Hook, NY, USA, 3323–3331. http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf

[15] Zach Harned and Hanna Wallach. 2019. Stretching Human Laws to Apply to Machines: The Dangers of a "Colorblind" Computer. *Florida State Univeristy Law Review* 47 (2019), 617.

[16] D. Hume, T.L. Beauchamp, P.P.S.R.S.T.L. Beauchamp, and Oxford University Press. 2000. *An Enquiry Concerning Human Understanding: A Critical Edition.* Clarendon Press, Oxford, UK.

[17] Heinrich Jiang and Ofir Nachum. 2020. Identifying and Correcting Label Bias in Machine Learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 108)*, Silvia Chiappa and Roberto Calandra (Eds.). PMLR, Online, 702–712. https://proceedings.mlr.press/v108/jiang20a.html

[18] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems. *CoRR* abs/1907.09615 (2019), 13. arXiv:1907.09615 http://arxiv.org/abs/1907.09615

[19] Nathan Kallus and Angela Zhou. 2018. Residual Unfairness in Fair Machine Learning from Prejudiced Data. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, Stockholm, Sweden, 2444–2453.

[20] F. Kamiran and T. Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.

[21] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Machine Learning and Knowledge Discovery in Databases*, Peter A. Flach, Tijl De Bie, and Nello Cristianini (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 35–50.

[22] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic Recourse: From Counterfactual Explanations to Interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 353–362. https://doi.org/10.1145/3442188.3445899

[23] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholm, Sweden, 2564–2572. http://proceedings.mlr.press/v80/kearns18a.html

[24] Thomas Kehrenberg, Myles Scott Bartlett, Oliver Thomas, and Novi Quadrianto. 2020. Null-sampling for Interpretable and Fair Representations. In *European Conference on Computer Vision – ECCV 2020*. Springer, Springer International Publishing, Glasgow, Scotland, 565–580.

[25] Thomas Kehrenberg, Zexun Chen, and Novi Quadrianto. 2020. Tuning Fairness by Balancing Target Labels. *Frontiers in Artificial Intelligence* 3 (2020), 33. https://doi.org/10.3389/frai.2020.00033

[26] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2019. Fairness through Causal Awareness: Learning Causal Latent-Variable Models for Biased Data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 349–358. https://doi.org/10.1145/3287560.3287564

[27] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. 2018. Learning Adversarially Fair and Transferable Representations. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, Stockholm, Sweden, 3381–3390. http://proceedings.mlr.press/v80/madras18a.html

[28] David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc., Montréal, Canada, 6147–6157. https://proceedings.neurips.cc/paper/2018/file/09d37c08f7b129e96277388757530c72-Paper.pdf

[29] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1 – 38. https://doi.org/10.1016/j.artint.2018.07.007

[30] Hussein Mozannar and David Sontag. 2020. Consistent Estimators for Learning to Defer to an Expert. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, Virtual, Online, 7076–7087. http://proceedings.mlr.press/v119/mozannar20b.html

[31] Song Park, Sanghyuk Chun, Junbum Cha, Bado Lee, and Hyunjung Shim. 2021. Multiple Heads are Better than One: Few-shot Font Generation with Multiple Localized Experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Online, 14.

[32] Judea Pearl. 2009. *Causality* (2 ed.). Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511803161

[33] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, Ying Li, Bing Liu, and Sunita Sarawagi (Eds.). ACM, Las Vegas, Nevada, USA, 560–568.

[34] Rare Recruitment. 2021. *Target Oxbridge*. Target Oxbridge. https://targetoxbridge.co.uk/the_programme.html

[35] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[36] Donald B Rubin. 1990. Formal mode of statistical inference for causal effects. *Journal of statistical planning and inference* 25, 3 (1990), 279–292.

[37] Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. 2018. Aequitas: A Bias and Fairness Audit Toolkit. *CoRR* abs/1811.05577 (2018), 19. arXiv:1811.05577 http://arxiv.org/abs/1811.05577

[38] Uri Shalit, Fredrik D. Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, Sydney, Australia, 3076–3085. http://proceedings.mlr.press/v70/shalit17a.html

[39] Viktoriia Sharmanska, Lisa Anne Hendricks, Trevor Darrell, and Novi Quadrianto. 2020. Contrastive Examples for Addressing the Tyranny of the Majority. *CoRR* abs/2004.06524 (2020), 16. arXiv:2004.06524 https://arxiv.org/abs/2004.06524

[40] Joshua Simons, Sophia Adams Bhatti, and Adrian Weller. 2021. Machine Learning and the Meaning of Equal Treatment. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 956–966. https://doi.org/10.1145/3461702.3462556

[41] Songül Tolan. 2019. Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges. *CoRR* abs/1901.04730 (2019), 22. arXiv:1901.04730 http://arxiv.org/abs/1901.04730

[42] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 10–19. https://doi.org/10.1145/3287560.3287566

[43] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review* 41 (2021), 105567. https://doi.org/10.1016/j.clsr.2021.105567

[44] Michael Wick, Swetasudha Panda, and Jean-Baptiste Tristan. 2019. Unlocking Fairness: a Trade-off Revisited. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., Vancouver, Canada, 8783–8792. http://papers.nips.cc/paper/9082-unlocking-fairness-a-trade-off-revisited.pdf

[45] Alice Xiang. 2021. Reconciling legal and technical approaches to algorithmic bias. *Tennessee Law Review* 88, 3 (2021), 75.

A RELATED WORKS

To the best of our knowledge, this is the first work to address positive action within the context of a decision support system. However, there are prior works that look at related problems. We briefly describe the most relevant ones to place the problem of determining positive action candidates in context.

Deferral. The challenge in learning to defer is identifying which candidates the model is uncertain about. Once identified, these candidates are referred to a human decision-maker which comes at some cost [28, 30]. This poses interesting questions about the practical quantification of uncertainty, and would be a potential extension to our framework. However, deferment differs from identifying positive action candidates, as the system we are adapting may be confident in it's assessment that a candidate who would be suitable for positive action should be rejected.

Actionable Recourse. Another related field is that of recourse. Works in this area, such as [18, 22, 42] aim to determine how the world would have had to be different for an alternative outcome to occur. They aim to explain what would need to change about a rejected candidate for them to be accepted. Our framework instead asks more direct questions: If a candidate were perceived to have an alternative protected attribute value, would the outcome be different? And, would the outcome change if the protected attribute and its causal consequences were different?

Auditing Systems. This is a multi-faceted, broad area, but in general auditing aims to evaluate either a dataset [37], or a system [23] for potential bias. Examples of auditing systems that are similar to ours include [5]. In their work the authors take an alternative counterfactual approach based on finding the nearest datapoint in the data with a different protected attribute and compare outcomes. These works differ in motivation as they use their auditing method to look at which *groups* are most affected, whereas we evaluate which *individuals* are likely to be affected.

B MODEL IMPLEMENTATION

The adversarial autoencoder has a similar architecture to [27], with multiple decoders [31, 38], and comprises:

(1) An encoder function $g : (\mathcal{X}, \mathcal{S}) \to \mathcal{Z}$ to map the input x to a more malleable representation z.
(2) An Adversary function $h : \mathcal{Z} \to \mathcal{S}$ to encourage the representation in the latent space to *not* be predictive of s.
(3) An ensemble of $\mathcal{S}$-specific decoders. The task is to produce a reconstruction x_s from z and is defined as a function $k : (\mathcal{Z}) \to \mathcal{X}_s \quad \forall\, s \in \mathcal{S}$. Where $\mathcal{X}_s$ is an array of reconstructions, each corresponding to a possible s-value. During training, $\mathcal{X}_s$ is indexed by the real s value so that only the $\mathcal{S}$-head that corresponds to the true protected attribute is used for training.

The encoder's purpose is to produce a *likely* counterfactual X with respect to S. To do this, we produce a latent embedding, Z which removes as much information about S as possible. Then, we have one decoder-head per possible S-label, allowing the effect of s to be reintroduced.[13] We train this model by optimising the objective function in Equation (1), where ℓ_{recon} is an appropriate loss between the reconstructions and the features, and ℓ_{adv} is the adversarial loss realised as cross-entropy between the predicted and target S coupled with a supplementary non-parametric measure

[13] This could be performed with a conditional decoder that additionally accepts the protected attribute as input, but in practice, we found our approach to work more consistently.

(Maximum Mean Discrepancy [13]) with a linear kernel between the embeddings per group (i.e. $\text{MMD}(Z_{s=0}, Z_{s=1})$). A hyper-parameter λ is incorporated to allow for a trade-off between the two competing losses.[14]

$$\mathcal{L}_{\text{AE}} = \min_{\theta,\pi} \max_{\phi} \mathbb{E}_{x\sim X}[\ell_{\text{recon}}(k_\pi(g_\theta(x), s)_s; x) - \lambda \ell_{\text{adv}}(h_\phi(g_\theta(x)), s)] \tag{1}$$

The classification model consists of a shared network with, in a similar fashion to the autoencoder, S-specific task-heads. This is to capture any potential direct discrimination that the model determines to exist based on past data. For the classification model the task is to produce an ensemble of predictions of the class label y_s from x and is defined as $f_s : (\mathcal{X}) \rightarrow \mathcal{Y}_s \quad \forall\, s \in \mathcal{S}$. As with the autoencoder, only the $\mathcal{S}$-head that corresponds to the true protected label is used for training. The objective is shown in the following equation:

$$\mathcal{L}_{\text{Clf}} = \min_{\omega,\xi} \mathbb{E}_{x\sim X}[\ell_{\text{pred}}(f_\omega(x)_s; y)] \tag{2}$$

At inference time, the autoencoder model produces one reconstruction per S-label, per sample, and likewise for the classification model. In the case of a binary S this produces two reconstructions per sample and two decisions per reconstruction, resulting in 4 outcomes per sample.

C EVALUATION METRICS

Acceptance percentage per colour (Acceptance). When this is equalised across groups, demographic parity is satisfied.

$$\text{Acceptance}(s) = P(Y = 1|S = s) \quad \forall s \in \mathcal{S} \tag{3}$$

With $Y = 1$ being the 'accepted' outcome.
True Capture percentage (TCP). This captures the rate of applicants with the ability to graduate, that are not rejected:

$$\text{TCP}(s) = P(Y \in \{1, 2\}|S = s, G = 1) \quad \forall s \in \mathcal{S} \tag{4}$$

With $Y = 2$ being the 'positive action candidate' outcome.
False Identification Difference (FID) measures the level of 'Equality of Opportunity' (EqOP), i.e. once a candidate is accepted, does their chance of graduating depend on the protected attribute? It is calculated as:

$$\text{FID} = |P(G = 0|S = 1, Y = 1) - P(G = 0|S = 0, Y = 1)| \tag{5}$$

Accuracy. We evaluate the utility of the model with regard to both Y, predicting a proxy-label based on the best assumptions from the data; and G, predicting the obscured 'true' outcome.

$$\text{Accuracy}(y) = P(\text{prediction} = y) \quad \forall y \in \mathcal{Y} \tag{6}$$

$$\text{Accuracy}(g) = P(\text{prediction} = g) \quad \forall g \in \mathcal{G} \tag{7}$$

D AUDITING THE UCI ADULT INCOME FOR BIAS

Figure 6 shows a counterfactual subgroup analysis for 3 protected attributes within the UCI Adult Income data set. We note that the accuracy of the PAF model is on par with baseline models. As before, the individuals within subgroups G_4 and G_3 did not achieve counterfactual consensus. For example, for sex, the subgroup G_4 contains males that are above the \$50, 000 threshold, but their female counterfactual counterparts would be under the threshold, whereas G_3 captures females under the threshold whose male counterfactual counterparts would be above the threshold. When comparing

[14] In our experiments, we use $\lambda = 1.0$

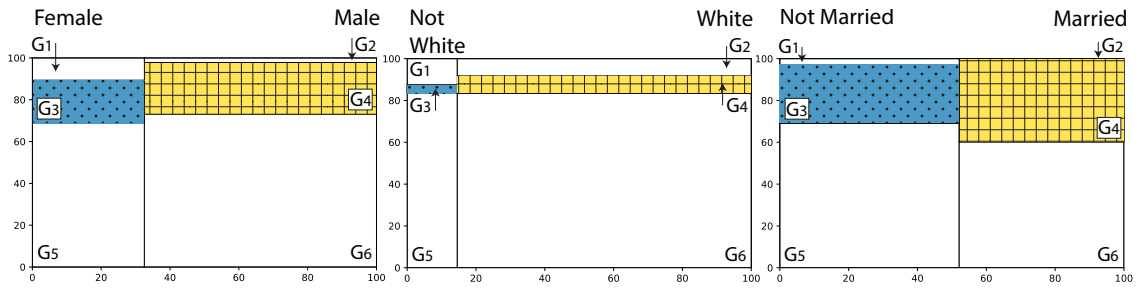

Fig. 6. Breakdown of group allocations on the withheld test set of the UCI Adult dataset averaged over 10 repeats, using 3 values as the protected attribute. **Left**: The binary 'sex' feature. **Middle**: The 'race' feature binarised to membership of the majority group (white). **Right**: The 'marital status' feature binarised to whether currently married. In all cases, the x-axis represents the percentage of the data that belongs to each protected attribute, while the y-axis represents the percentage of the population assigned each outcome. Group membership is defined in Table 1. For all attributes, subgroups G_3 and G_4 highlight the proportion of the population for which intervening on the protected attribute will result in the outcome changing as well. In G_3 the outcome changes from negative to positive when S changes, while changes in G_4 result in the opposite outcome. Although an intriguing visualisation of the effect of the different attributes, conclusions should be drawn carefully as the attributes can act as a proxy to hidden patterns in the data. Further discussion can be found in Section D.

the effects of the 3 protected attributes we examined, we can see that changing the marital status is most likely to result in a change the outcome. That alone, however, is not enough to deduce a causal relationship between marital status and salary. Martial status might have a direct impact on salary but it is also a proxy to other relevant attributes such as age, for example. Similarly, the effect of sex can be a combination of a direct effect and a proxy effect from additional relevant attributes such as occupation type and working hours. The effect of race may seem to be the least influential of the 3, but this can be misleading as we can't quantify the contributions of proxy effects on marital status and sex.

Employing our approach on the adult dataset highlights some important challenges and choices: Protected attributes may be correlated to other attributes that are relevant to the task. These can add proxy effects and mask the direct effect of the protected attribute. We can choose to leave the re-calibration of features unrestricted or to keep some features as they were originally. The latter may impact the quality of counterfactual representations we can achieve [33].

There can be some disparity in the opposite direction to what we expect. When comparing males to females in the adult dataset, overall, the direction of bias is in favour of males. We do detect, however, females earning above the $\$50,000$ threshold, whose male counterparts would be under the threshold. This 'reversed' bias could be present in a subset of occupations.

Under-represented groups can be separated from the majority by a combination of protected attributes. Analysing each protected attribute separately does not capture any compounding effects that might be experienced by a specific under-represented group, for example, unmarried, not-white women [6, 10]. Considering every possible combination, however, will significantly increase the size of the feature space and as a result, the required size of training data.

E FULL RESULTS

A measurement of the plausibility of the reconstructed counterfactuals can be found in Table 6. For transparency we report two versions of our model `PAF Y={1}` and `PAF Y={1,2}`. These correspond to how we consider positive action candidates in the metrics. In `PAF Y={1,2}` positive action candidates are reported as receiving a successful (positive) outcome. For comparison, in the `PAF Y={1}` positive action candidates are reported as receiving a negative outcome.

Table 4. Full results on the UCI Adult Income Dataset.

Protected Attribute	Metric	LR	FairLearn	K & C	PAF Y={1} Ours	PAF Y={1,2} Ours
Binary Married	Accuracy	84.77 ± 0.40	83.14 ± 0.39	81.88 ± 0.51	85.17 ± 0.41	73.10 ± 2.47
	DP Diff	37.15 ± 0.86	19.03 ± 0.98	13.28 ± 0.74	36.89 ± 4.44	8.85 ± 4.22
	TPR Diff	37.15 ± 0.86	7.52 ± 3.71	16.71 ± 3.40	32.58 ± 3.60	4.51 ± 2.86
Binary Race	Accuracy	84.85 ± 0.44	84.66 ± 0.42	84.67 ± 0.40	85.20 ± 0.43	84.96 ± 0.48
	DP Diff	8.93 ± 0.76	8.66 ± 0.74	8.40 ± 0.63	9.52 ± 1.43	5.42 ± 1.24
	TPR Diff	4.32 ± 3.25	4.61 ± 2.86	3.87 ± 2.77	5.77 ± 3.09	3.13 ± 1.64
Binary Sex	Accuracy	84.77 ± 0.45	84.43 ± 0.49	84.09 ± 0.45	84.80 ± 0.50	78.78 ± 1.73
	DP Diff	17.57 ± 0.74	14.53 ± 0.94	9.15 ± 0.77	16.31 ± 1.96	4.40 ± 3.28
	TPR Diff	7.23 ± 2.41	2.27 ± 1.87	12.95 ± 3.19	2.27 ± 1.92	7.94 ± 2.44

Table 5. Full Breakdown of results. All probability-based metric results are converted to percentages. $S = 0$ corresponds to green applicants in the example from the main text, and $S = 1$ corresponds to blue applicants.

Metric	Data	DemPar	LR	PAF Y={1} (Ours)	K&C	Kamishima	FairLearn	PAF Y={1,2} (Ours)
Acc. Y ↑	—	84.50 ± 0.43	98.51 ± 0.18	98.31 ± 0.26	97.27 ± 0.27	97.37 ± 0.31	92.64 ± 0.41	84.54 ± 2.06
Acc. G ↑	—	79.25 ± 9.48	87.05 ± 0.46	86.67 ± 0.47	88.40 ± 0.54	88.66 ± 0.47	86.16 ± 0.44	73.63 ± 2.26
DP Diff ↓	31.43 ± 0.88	0.01 ± 0.01	29.09 ± 1.09	29.11 ± 1.32	25.98 ± 1.13	31.37 ± 0.77	18.81 ± 1.45	3.62 ± 3.00
TPR Diff ↓	—	35.30 ± 45.57	4.28 ± 0.48	3.11 ± 1.46	11.34 ± 0.84	50.46 ± 5.85	13.01 ± 6.28	3.11 ± 1.46
FID ↓	4.19 ± 3.56	23.14 ± 14.59	3.73 ± 2.94	4.16 ± 3.04	8.97 ± 4.66	15.74 ± 7.89	20.74 ± 2.95	31.91 ± 1.84
$P(Y = 1\|S = 1)$	35.73 ± 0.77	23.12 ± 16.31	34.52 ± 0.79	35.15 ± 1.36	31.68 ± 0.83	33.18 ± 0.84	26.65 ± 1.17	35.15 ± 1.36
$P(Y = 1\|S = 0)$	4.30 ± 0.33	23.13 ± 16.31	5.43 ± 0.52	6.04 ± 0.60	5.70 ± 0.53	1.80 ± 0.26	7.83 ± 0.52	33.78 ± 4.85
$P(G = 1\|S = 1)$	15.48 ± 0.76	15.48 ± 0.76	15.48 ± 0.76	15.48 ± 0.76	15.48 ± 0.76	15.48 ± 0.76	15.48 ± 0.76	15.48 ± 0.76
$P(G = 1\|S = 0)$	3.07 ± 0.42	3.07 ± 0.42	3.07 ± 0.42	3.07 ± 0.42	3.07 ± 0.42	3.07 ± 0.42	3.07 ± 0.42	3.07 ± 0.42
$P(Y = 1\|S = 1, G = 1)$	93.82 ± 1.36	60.57 ± 42.80	91.70 ± 1.92	92.63 ± 2.08	92.51 ± 1.72	93.19 ± 1.37	71.51 ± 3.72	92.63 ± 2.08
$P(Y = 1\|S = 0, G = 1)$	60.38 ± 4.73	69.83 ± 6.03	69.84 ± 5.00	73.07 ± 4.94	67.59 ± 5.66	33.92 ± 5.29	53.17 ± 6.02	96.74 ± 1.93
$P(Y = 0\|S = 1, G = 0)$	74.92 ± 0.55	83.77 ± 11.41	75.96 ± 0.62	75.38 ± 0.96	79.46 ± 0.74	77.82 ± 0.70	81.58 ± 0.84	75.38 ± 0.96
$P(Y = 0\|S = 0, G = 0)$	97.48 ± 0.30	78.34 ± 16.59	96.61 ± 0.44	96.08 ± 0.48	96.26 ± 0.49	99.21 ± 0.25	93.60 ± 0.52	68.21 ± 4.80
$P(Y = 1\|S = 1, G = 0)$	25.08 ± 0.55	16.23 ± 11.41	24.04 ± 0.62	24.62 ± 0.96	20.54 ± 0.74	22.18 ± 0.70	18.42 ± 0.84	24.62 ± 0.96
$P(Y = 1\|S = 0, G = 0)$	2.52 ± 0.30	21.66 ± 16.59	3.39 ± 0.44	3.92 ± 0.48	3.74 ± 0.49	0.79 ± 0.25	6.40 ± 0.52	31.79 ± 4.80
$P(G = 1\|S = 1, Y = 1)$	40.64 ± 1.78	40.25 ± 3.42	41.11 ± 1.96	40.78 ± 1.50	45.19 ± 2.07	43.47 ± 1.83	41.53 ± 2.47	40.78 ± 1.50
$P(G = 1\|S = 0, Y = 1)$	43.14 ± 5.19	21.64 ± 19.56	39.56 ± 4.88	37.15 ± 3.87	36.48 ± 4.76	58.20 ± 10.21	20.79 ± 2.72	8.87 ± 1.08
$P(G = 0\|S = 1, Y = 1)$	59.36 ± 1.78	59.65 ± 3.42	58.89 ± 1.96	59.22 ± 1.50	54.81 ± 2.07	56.53 ± 1.83	58.47 ± 2.47	59.22 ± 1.50
$P(G = 0\|S = 0, Y = 1)$	56.86 ± 5.19	78.36 ± 19.56	60.44 ± 4.88	62.85 ± 3.87	63.52 ± 4.76	41.80 ± 10.21	79.21 ± 2.72	91.13 ± 1.08
$P(G = 1\|S = 1, Y = 0)$	1.48 ± 0.27	6.62 ± 6.62	1.95 ± 0.39	1.74 ± 0.42	1.69 ± 0.35	1.57 ± 0.30	5.99 ± 0.63	1.74 ± 0.42
$P(G = 1\|S = 0, Y = 0)$	1.27 ± 0.23	1.22 ± 0.22	0.98 ± 0.20	0.88 ± 0.21	1.06 ± 0.23	2.07 ± 0.36	1.57 ± 0.36	0.14 ± 0.08
$P(G = 0\|S = 1, Y = 0)$	98.52 ± 0.27	93.28 ± 6.62	98.05 ± 0.39	98.26 ± 0.42	98.31 ± 0.35	98.43 ± 0.30	94.01 ± 0.63	98.26 ± 0.42
$P(G = 0\|S = 0, Y = 0)$	98.73 ± 0.23	98.78 ± 0.22	99.02 ± 0.20	99.12 ± 0.21	98.94 ± 0.23	97.93 ± 0.36	98.43 ± 0.36	99.86 ± 0.08

A full breakdown of results can be found in Table 5. Further results on the UCI Adult Income Dataset can be found in Table 4.

F DATA GENERATION

We first draw samples for S from a Bernoulli distribution, and model the underlying construct as a Uniform distribution (Figure 7(i)) — this is where the WAE worldview is applied, as $\tilde{X}_{apt}$ is independent of S: $S \sim \mathcal{B}(0.5)$ and $\tilde{X}_{apt} \sim \mathcal{U}(0, 1)$. To represent the imbalance of opportunity between the groups, for example, due to variation in parental support between blue and green parents, we map from the uniform distribution to an S-conditioned distribution for each feature using an inverse-CDF (percent point) function, $\tilde{X}_{apt}$ to $\tilde{X}$. This mapping is captured by $\Delta\tilde{X}_{s=0,1}$ (Figure 7(ii)).

The features $\tilde{X}_{s=0,1}$ are still in the construct space, representing the potential to successfully graduate from the university course at the point of applying. The mapping between $\tilde{X}$ and X is made of two noisy observations for each

Table 6. Reconstruction performance on withheld data. Ideally the error rate on the reconstructions should not be worse than the reconstruction on the ground truth.

Feature	L_1 Reconstruction	L_1 Reconstruction Counterfactual
1	0.0071 ± 0.0039	0.0085 ± 0.0057
2	0.0074 ± 0.0038	0.0073 ± 0.0043
3	0.0078 ± 0.0076	0.0305 ± 0.0254
4	0.0305 ± 0.0254	0.0123 ± 0.0071
5	0.0123 ± 0.0071	0.0266 ± 0.0143
6	0.0090 ± 0.0064	0.0088 ± 0.0063

Table 7. Percentage of the data assigned to each group in the UCI Adult income dataset.

Group	S="Sex Male"	S="Race White"	S="Married"
1	0.79 ± 0.18	0.81 ± 0.14	1.03 ± 0.28
2	1.44 ± 0.67	6.61 ± 2.24	0.18 ± 0.21
3	6.70 ± 1.45	0.57 ± 0.15	14.61 ± 2.66
4	16.77 ± 1.75	6.84 ± 0.78	18.86 ± 2.49
5	20.71 ± 1.65	10.54 ± 0.51	35.65 ± 2.99
6	44.69 ± 2.89	55.85 ± 2.38	28.42 ± 2.37
N/A	8.89 ± 1.33	18.78 ± 1.35	1.26 ± 0.85

Table 8. Breakdown of the $G1$ group, comprising individuals funnelled into this group due to different selection rules. Consensus corresponds to selection rule 1 in Table 1. Direct bias corresponds to selection rules 2 and 8 in Table 1. Fallback indicates bias was detected in the opposite direction and the decision reverted to the original outcome.

Selection Rule	Sex	Race	Marital Status
Consensus	4.3	69.2	71.7
Direct bias	18.4	0	0
Fallback	77.2	30.8	28.3

Table 9. Results demonstrating the result of a post-hoc Logistic Regression (LR) classifier trained with 5-fold cross validation to predict S at various points in our model – ideally the Z space should be no more predictive than the `Majority` classifier model. In addition, we also show the performance of both our classifier model and a post-hoc baseline predicting the classification target from the reconstructions. Ideally, these numbers should be similar to each other, demonstrating that predictive power has been retained, despite not being explicitly optimised for.

Model	Input	Target	Accuracy
LR	Enc Z	S	61.65 ± 7.07
LR	Recon X	S	85.79 ± 3.94
LR	X	S	82.81 ± 0.70

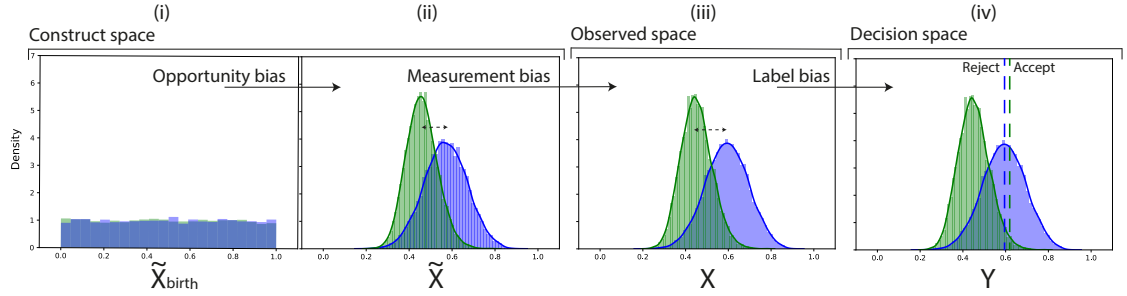

Fig. 7. Changes in the engineered synthetic data. Starting from a uniform distribution, we visualise how the additive effect of bias can result in a significant disproportion of success between groups differing by a protected attribute. The opportunity bias and measurement bias are modelled as a shift between the distributions. The label bias is modelled by having different acceptance thresholds for the different groups (vertical dashed lines in the right figure).

feature. A measurement bias further aggravates the disparity between the blue and green distributions (Figure 7(iii)). We then generate two outcome scores: 1. An 'acceptance score' based on a linear combination of the observed features. When mapping from X to Y, from the observed to the decision space, we add a label bias by setting different acceptance

thresholds depending on the value of S (Figure 7(iv)). 2. A 'graduation grade' based on a linear combination of the features in $\tilde{X}$, bypassing the effect of the introduced measurement bias and label bias.

We first draw samples for S from a Bernoulli distribution ($\mathcal{B}$), and model the underlying construct as a Uniform distribution ($\mathcal{U}$) such that $\tilde{X}_b$ is independent of S:

$$S \sim \mathcal{B}(0.5) \quad \text{and} \quad \tilde{X}_b \sim \mathcal{U}(0,1)\ .$$

We then take the inverse-CDF (product point function) of a distribution at point $\tilde{x}_b$ ($CDF^{-1}(\text{distribution}, \text{point})$) for three unobserved features.

$$\tilde{x}_0 \sim \begin{cases} CDF^{-1}(\mathcal{N}(0.65, 0.15), \tilde{x}_b), & \text{if } s = 1 \\ CDF^{-1}(\mathcal{J}_U(-2, 3, 0.35, 0.2), \tilde{x}_b), & \text{otherwise} \end{cases}$$

$$\tilde{x}_1 \sim \mathcal{N}(0.4 + (2s - 1), 0.2)$$

$$\tilde{x}_2 \sim \begin{cases} CDF^{-1}(\mathcal{L}(0.5, 0.075), \tilde{x}_b), & \text{if } s = 1 \\ CDF^{-1}(\mathcal{T}(100, 0.4, 0.15), \tilde{x}_b), & \text{otherwise} \end{cases}$$

Where $\mathcal{N}$ is a Normal distribution, $\mathcal{J}_U$ is Johnsons-SU distribution, $\mathcal{L}$ is a Laplace distribution and $\mathcal{T}$ is a Student-T distribution.

We then have two observers of each feature. Both observers add noise from a Normal distribution, but with different mean and standard deviation.

$$\begin{aligned} \tilde{x}_0 \text{ Observer 1}: &\quad \tilde{x}_0 + \mathcal{N}(0.03, 0.02) \\ \tilde{x}_0 \text{ Observer 2}: &\quad \tilde{x}_0 + \mathcal{N}(0.01, 0.04) \\ \tilde{x}_1 \text{ Observer 1}: &\quad \tilde{x}_1 + \mathcal{N}(0, 0.02) \\ \tilde{x}_1 \text{ Observer 2}: &\quad \tilde{x}_1 + \mathcal{N}(0, 0.05) \\ \tilde{x}_2 \text{ Observer 1}: &\quad \tilde{x}_2 + \mathcal{N}(0.03, 0.01) \\ \tilde{x}_2 \text{ Observer 2}: &\quad \tilde{x}_2 + \mathcal{N}(0.01, 0.02) \end{aligned}$$

The admittance score (Y) is based on a combination of the mean observation per feature. Let N be the number of observers.

$$\begin{aligned} \tilde{Y} = &0.4\left(\frac{1}{N}\sum_{i=0}^{N} \tilde{x}_0 \text{ Observer }_i\right) + \\ &0.4\left(\frac{1}{N}\sum_{i=0}^{N} \tilde{x}_1 \text{ Observer }_i\right) + \\ &0.2\left(\frac{1}{N}\sum_{i=0}^{N} \tilde{x}_2 \text{ Observer }_i\right) \end{aligned}$$

Then, to incorporate direct discrimination, a factor γ is added to the admittance score.

$$Y = \begin{cases} \tilde{Y} + \gamma, & \text{if } s = 1 \\ \tilde{Y} - \gamma, & \text{otherwise} \end{cases}$$

During our experiments we set $\gamma = 0.01$.

We also model the final graduation grade. We model this as a binary label, "good graduating grade" or "not good graduating grade". This is based on the unobserved score for each feature, and is different per subgroup to reflect that one measure of success need not be consistent across all of the population.

$$G = \begin{cases} 0.3\tilde{x}_0 + 0.25\tilde{x}_1 + 0.45\tilde{x}_2, & \text{if } s = 1 \\ 0.1\tilde{x}_0 + 0.7\tilde{x}_1 + 0.2\tilde{x}_2, & \text{otherwise} \end{cases}$$